The Philosophical Quarterly Vol. 00, No. 0 *2022*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqab069

ON THE LOGICALITY OF TRUTH

By Kentaro Fujimoto

Deflationism about truth describes truth as a logical notion. In the present paper, I explore the implication of the alleged logicality of truth from the perspective of axiomatic theories of truth, and argue that the deflationist doctrine of the logicality of truth gives rise to two types of self-undermining arguments against deflationism, which I call the conservativeness argument from logicality and the topic-neutrality argument.

Keywords: truth, deflationism about truth, the conservativeness argument, axiomatic theories of truth, the logicality of truth.

I. INTRODUCTION

Deflationism about truth is often associated with the claim that truth is a logical notion.[1] However, the exact sense in which truth is logical is often left unexplained or unclear in the literature. The logicality of truth is often propounded along with the claim that the purpose of truth is to serve as a linguistic device to increase our expressive power.[2] Deflationists typically see truth as a device of generalisation via its function to express infinite conjunctions and perform indirect endorsements. With this understanding of truth, one may well draw an analogy between first-order quantification and truth: the former is a device

[1] Field (1994, 1999) and Horwich (1998, 2010) are, perhaps, the two most influential deflationists who describe truth as a logical notion; other examples are Hill (2002), Damnjanovic (2005), Bonnay and Galinon (2018), and McGinn (2000) (though McGinn denies that he is a deflationist). Nowadays, the logicality claim of truth is often considered and discussed as a core tenet of deflationism; see also Ketland (1999), Bar-On et al. (2000), Damnjanovic (2010), and Wyatt (2016). Some authors have certain reservations in calling truth simply logical; Künne (2003) calls truth 'broadly logical', Horsten (2011) takes truth as a 'logico-linguistic' notion, and Leitgeb (2007) and Picollo and Schindler (2018) call it 'quasi-logical'. Nonetheless, they make such qualifications only because truth presupposes its bearers, which are putatively non-logical entities, and, seemingly, still regard truth as on a par with other standard logical notions under the presupposition of these bearers; see also Section II.

[2] See Horwich (1998), Field (1999), Damnjanovic (2010), and Horsten (2011), for example.

of generalisation over objects and commonly considered logical, and the latter is a device of generalisation over sentences (or their referents). This analogy understandably explains deflationists' tendency to call truth logical. However, it still remains to be articulated why truth is logical by virtue of being such a device, and a mere analogy falls short of a satisfactory answer to this question.

One central feature of logic is topic-neutrality. Anything logical should be applicable to any subject and topic in the same uniform way. Hence, if truth is a logical notion, then it ought to be topic-neutral in this sense, namely, that it is universally applicable to every subject in a uniform way that does not depend on the choice of subject and its theorisation. This is the assumption with which I start. The deflationist conception of truth as a linguistic tool for an expressive purpose of the aforementioned kind especially fits this intuition: an infinite conjunction can be formed from any sentences about any subject matter, any sentence about any subject matter can be indirectly endorsed, and so on. Hence, it seems a natural working hypothesis that it is this topic-neutrality that prompts deflationists to call truth logical, and even if one attributes logicality to truth for a different reason,[3] the alleged logicality of truth requires its topic-neutrality anyway. I thus formulate the deflationist tenet of the logicality of truth into the following thesis, with a particular emphasis on the topic-neutral aspect of logic.

Logicality thesis: Truth is a logical linguistic device that is universally applicable to any subject and its theorisation in the same uniform way.

The aim of the present paper is to explore the philosophical implications of this logicality thesis from the perspective of formal theories of truth and thereby present a new type of argument that challenges deflationism.

II. PRELIMINARY FORMAL ASSUMPTIONS

How should truth be implemented in formal theories of truth? There are different options, and I will focus on one particular, but probably nowadays the customary and most popular, formal setting and process in which truth is formally implemented.

Firstly, I will exclusively focus on the notion of truth as a predicate. In other words, the present paper is engaged in the study of the predicate 'is true'. There are other ways of conceiving of truth, such as prosententialism, but I will not consider them in the present paper.

[3] For example, Hill (2002), Künne (2003), and McGinn (2000) seem to treat truth as a logical notion primarily for different reasons: Hill and Künne both contend that truth is reducible to other logical (or 'broadly logical') notions, and McGinn appeals to the simplicity and primitiveness of truth and the fundamental role in our thought that truth plays (though he seems to think that truth plays such a role exactly by virtue of the expressive power that truth brings to us).

Secondly, I adopt the so-called axiomatic approach to truth, in which theories of truth are formulated as recursively axiomatisable first-order theories and the ordinary recursive (effective) notion of first-order logical consequence is employed.[4] Throughout the present paper, by a 'theory' I always mean a recursive set of axioms in the language in question, and, given two theories B and C in the same language, we will write $\mathsf{C} \vdash \mathsf{B}$ when every axiom of B is provable in C.

Thirdly, by a 'theory of truth', I mean a theorisation of the discourse of truth and falsehood about some non-semantic subject. Namely, I will always presuppose some non-semantic subject of discourse, such as arithmetic and physics, and focus on theories of truth about the subject, which I will call the *base subject* of the theories of truth. Accordingly, a theory of truth, as a whole, will be always assumed to embody a certain theorisation of its base subject.

Fourthly, I assume that the bearers of truth are sentence types.[5] Since we have stipulated that truth is represented by a predicate of a first-order language, a theory of truth should be able to treat sentence types as first-order objects and embody a first-order theorisation of them. Hence, every theory of truth should comprise a sufficient theory of syntax as the theory of the bearers of truth. There are, however, different methods of incorporating a theory of syntax into a theory of truth. In the present paper, I will mainly focus on the customary one adopted by the majority of the theories of truth proposed thus far: namely, I will assume that the theorisation of the base subject in a theory of truth includes such a theory of syntax (*per se* or via coding). I will also consider an alternative setting later in Section IV in which the base subject and syntax are separately theorised within a theory of truth.

Lastly, following the convention, I will mainly consider arithmetic as the base subject of theories of truth. It seems almost unanimously agreed that a sufficient part of arithmetic for encoding a sufficient theory of syntax for theories of truth can be theorised without appealing to the notion of truth and thus by means of a theory in the first-order language $\mathcal{L}_{\mathbb{N}}$ of arithmetic without the truth predicate T. Hence, the last (fourth) requirement is met by

[4] There is a number of alternatives. The semantic (or model-theoretic) approach to truth aims to define an extension of a truth predicate on a given fixed model-theoretic structure. However, on the one hand, focusing on one specific structure seems unsuitable for the study of the logical aspect of truth, and, on the other hand, if we study a semantic theory of truth on all arbitrary relevant first-order structures, then the study boils down to the axiomatic approach after all, because of the completeness theorem, as long as we treat truth as a first-order predicate. Another example is to adopt a strong non-effective logic; Shapiro (1998) suggests this approach as a consequence of his conservativeness argument.

[5] This assumption is not indispensable for my argument. My argument can be applied to any other setting in which the bearers of truth have a sufficiently similar structure to that of sentence types and are equipped with operations corresponding to the first-order logical operations, such as negation, quantification, predicate application, etc; see Halbach (2010, section 2) for more discussion.

any theory of truth that includes such an $\mathcal{L}_{\mathbb{N}}$-theory. Let $\mathcal{L}_{\mathbb{N}}^{+}$ be the language $\mathcal{L}_{\mathbb{N}}$ of arithmetic augmented with T. Accordingly, theories of truth that I will consider in the present paper are recursive theories in the language $\mathcal{L}_{\mathbb{N}}^{+}$ unless otherwise specified.

The fourth stipulation entails that a theory of truth commits us to some (putatively) non-logical entities and theory, namely, syntactic objects (or their codes) and a theory of syntax. Indeed, Halbach (2001) observed that a barely minimal condition for a theory of truth necessitates the existence of at least two objects. However, when deflationists say that truth is logical, they do not (and should not) mean logicality in such a strong sense that truth incurs absolutely no ontological commitment and presupposes no non-logical theory. Truth necessitates its bearers, and a theorisation of truth calls for a theorisation of the bearers. This is exactly why some philosophers call truth 'broadly logical', 'logico-syntactic', 'quasi-logical', and the like, instead of simply calling it logical (see fn 1). Hence, what the logicality thesis requires is that truth should function as a logical device in essentially the same or equivalent way as other logical devices, such as first-order quantification and conjunction, *under the presupposition of syntactic objects*, and that an adequate theory of truth should formally capture such a function of truth *under the presupposition of, and possibly in collaboration with, an adequate theory of syntax*.

III. THE CONSERVATIVENESS ARGUMENT

We will first see that the logicality thesis provides a new basis for an existing argument against deflationism about truth, namely, the so-called *conservativeness argument*, which was presented independently by Horsten (1995), Shapiro (1998), and Ketland (1999), and is probably the most influential argument against deflationism from the point of view of formal logic.

III.1 The orthodox argument

The conservativeness argument originally aims to challenge not the logicality thesis, but another core tenet of deflationism, which I call the *insubstantiality thesis*.

Insubstantiality thesis: Truth is a metaphysically and/or epistemologically insubstantial property with no explanatory power.

This formulation begs further clarification of what it means that truth is metaphysically or epistemologically insubstantial and that it has no explanatory power; the answer indeed varies among deflationists (and their opponents), and a lively debate is still ongoing regarding what it should, and should not,

mean.[6] Nevertheless, no matter what it is taken to mean, the conservativeness argument only needs to construe the insubstantiality thesis as necessitating the following requirement.

Conservativeness requirement: For every $\mathcal{L}_{\mathbb{N}}$-theory B that can encode a sufficient theory of syntax for theories of truth, the result $\mathcal{T} + \mathsf{B}$ of adding a set $\mathcal{T}$ of $\mathcal{L}_{\mathbb{N}}^{+}$-sentences, as axioms of truth, to B should be conservative over B, in the sense that $\mathcal{T} + \mathsf{B}$ proves exactly the same $\mathcal{L}_{\mathbb{N}}$-theorems as B does, if $\mathcal{T}$ only comprises axioms essential to truth.

The rationale for this requirement is that for any given *truth-free* theorisation of the subject of discourse, the addition of truth should not bear any substantial impact on the theorisation: if the conservativeness requirement is not met, then truth would be considered as making a substantial contribution to one's knowledge-gathering and explanation about the subject and attributing a new metaphysical or modal status to some statements purely about the subject, by virtue of its essential nature and function.[7] Given a theory of truth of the form $\mathcal{T} + \mathsf{B}$, we will call B the *base theory* of the theory, and, as we stipulated in Section II, we always assume that a base theory is rich enough to encode a sufficient theory of syntax for theories of truth.

Note that the conditional clause in the conservativeness requirement, namely, that $\mathcal{T}$ consists of axioms *essential to truth*, cannot be dropped. For, as Field (1999) points out, if $\mathcal{T}$ contains axioms that are not essential to truth and not postulated by virtue of the deflationist concept of truth, then the failure of conservation could be disregarded by deflationists as a result not of the deflationist concept of truth, but of something else. Indeed, many theories of truth on the market may be regarded as containing axioms involving T but not essential to truth. For example, many standard theories of truth of set theory contain the axiom scheme of replacement extended for the formulas contain-

[6] For example, Bar-On et al. (2000), Bar-On and Simmons (2007), Damnjanovic (2005), and Edwards (2013) give interesting discussions of the insubstantiality thesis from the perspectives of truth-conditional semantics, the illocutionary function of truth, the causal-explanatory power of truth, and the metaphysical distinction of abundant and sparse properties, respectively. See Damnjanovic (2010) and Wyatt (2016) for more extensive overviews of the debate.

[7] The conservativeness requirement was originally posed by *opponents* of deflationism, namely, Horsten (1995), Shapiro (1998), and Ketland (1999), but some self-acknowledged deflationists, such as Field (1999), accept it as well. The conservativeness requirement is nowadays one of the central topics in the debate on deflationism, and various arguments for and against it have been presented; for example, on the one hand, Cieśliński (2015) gives a meticulous argument against the association of the conservativeness requirement with the deflationist tenet of the insubstantiality of truth, and Picollo and Schindler (2018) argue that non-conservativeness is rather an expected and desirable phenomenon for deflationism; on the other hand, Fischer (2015) defended the conservativeness requirement from an instrumentalist understanding of deflationary truth, and Strollo (2013) contends that an even stronger type of conservativeness, the so-called semantic conservativeness, is required of deflationism; see also Halbach (2010), Horsten (2011), and Waxman (2017).

ing T.[8] Instances of this scheme containing T are natural axioms for theories of truth of set theory, but one may well regard them as not essential to truth because they are postulated by virtue of, say, the standard understanding of the axiom scheme of replacement as 'indefinitely extensible' and of the standard conception of the set-theoretic universe as closed under functions, and also because one would not postulate them when the base subject is changed to other subjects, such as arithmetic.

Now, the general structure of the conservativeness argument is as follows:

(C1) the conservativeness requirement must be met;
(C2) there are some necessary condition for an adequate theory of truth and some (standard) base theory B, such as PA, such that every theory of truth with the base theory B that meets the condition is not conservative over B;
(C3) therefore, the insubstantiality thesis and thus deflationism about truth are untenable.

The conservativeness argument depends on the necessary condition for an adequate theory of truth that one employs in the premise (C2), and several different versions of the conservativeness argument have been so far presented.

Let us first see a paradigmatic example due to Shapiro (1998) and Ketland (1999). We write $Prv_{\mathsf{B}}(x)$ for a canonical predicate expressing the provability in an $\mathcal{L}_{\mathbb{N}}$-theory B. The *Global Reflection Principle* for B, henceforth *GRef*(B), is defined as $\forall x(Prv_{\mathsf{B}}(x) \rightarrow \mathrm{T}x)$, which expresses that every theorem of B is true. Shapiro (1998) takes the provability of *GRef*(B) as a necessary condition for an adequate theory of truth for the following reason:

> [O]ne can state in $[\mathcal{L}_{\mathbb{N}}^{+}]$ that all of the axioms of $[\mathsf{B}]$ are true and one can state in $[\mathcal{L}_{\mathbb{N}}^{+}]$ that the rules of inference preserve truth. Since these generalizations are obviously correct, an adequate theory of truth should have the resources to *establish* them. It follows, or should follow, that all of the theorems are true. (Shapiro , 1998, 498)

Ketland (1999) also takes the provability of *GRef*(B) as a necessary condition for an adequate theory of truth for a similar reason:

> Part of the basic (not necessarily deflationist) idea about truth is that a particular statement φ and its "truth" . . . are somehow "equivalent". . . . But we must go further. Any *adequate* theory of truth should be able to prove the "equivalence" of a (possibly infinitely axiomatized) *theory* $[\mathsf{B}]$ and its "truth" [*GRef*(B)] . . . . (Ketland 1999, 90)

[8] See Fujimoto (2012) for several examples of such theories.

Here, for Ketland, to 'prove the "equivalence" of a theory [B] and its "truth"' means to prove *GRef*(B). Following Ketland's (1999, 87) terminology, let us call this adequacy condition the *generalised equivalence principle*[9]:

Generalised equivalence principle (*GEP*): An adequate theory of truth with the base theory B ought to prove *GRef*(B).

In addition to *GEP*, we also consider the following condition.

Disquotationality requirement (*Disq*): An adequate theory of truth with any base theory ought to prove every instance of the T-schema restricted to $\mathcal{L}_{\mathbb{N}}$, namely, $\mathrm{T}\ulcorner\sigma\urcorner \leftrightarrow \sigma$ for every $\mathcal{L}_{\mathbb{N}}$-sentence σ; we will denote the set of all these instances by $\mathcal{TB}$.

This condition *Disq* is nearly unanimously accepted as a necessary condition for an adequate theory of truth. Finally, Shapiro (1998) and Ketland (1999) thereby conclude that there is no adequate deflationist theory of truth with any arithmetical base theory B, since *GEP* and *Disq* jointly necessitate the provability of the consistency statement *Con*(B) for B, which conflicts with the conservativeness requirement because of Gödel's second incompleteness theorem.

III.2 The conservativeness argument from logicality

The conservativeness requirement is a central premise of the conservativeness argument. It has traditionally been advocated on the basis of the insubstantiality thesis, more or less in the way I described in Section III.1, in which the logicality thesis is not utilised. Let us call the conservativeness argument that employs the insubstantiality thesis in justification of the conservativeness requirement the *conservativeness argument from insubstantiality*. In the present subsection, I will present an argument for the conservativeness requirement on the basis of the logicality thesis, instead of the insubstantiality thesis.

The logicality thesis (partly) asserts that truth is topic-neutral and independent of the subject of discourse and its theorisation. Whilst we can talk of truths of any different subjects as separate issues on their own rights, such as the truth of arithmetic and that of physics, they are regarded as applications, or restrictions, of a more general, universal, notion of truth to the particular subjects. Hence, the essential nature of truth should be characterised independently of, and invariantly across, different subjects and their different theorisations.

With this 'logical' conception of truth, we naturally require a certain set $\mathcal{E}$ of 'logical' axioms of truth. Such an $\mathcal{E}$ should be compared to the first-order logical axioms of logical connectives and quantifiers, which are independent of

[9] Leitgeb (2007) also considers *GEP* as one of the desidarata for theories of truth.

the non-logical axioms that one postulates in one's theorisation of one's subject of discourse, and depend only on the purely syntactic feature of the language employed in the theorisation. Similarly, $\mathcal{E}$ should only (but adequately) theorise the essential 'logical' nature and function of truth, invariantly and uniformly across all subjects and their theorisations modulo a language.[10] Hence, truth should be applicable, in the same uniform way, to any given truth-free theorisation B of a non-semantic subject of discourse that meets the minimal requirement laid out in Section II, namely, that B includes a sufficient theory of syntax, and this process of applying truth to B results in a new theory $\mathcal{E} + \mathsf{B}$ in the language obtained by augmenting the language of B with the truth predicate T. When a theory $\mathcal{E} + \mathsf{B}$ of truth is obtained in such a way, we call B the *base theory* of $\mathcal{E} + \mathsf{B}$ as before.

We have particularly chosen arithmetic as our base subject and taken for granted (in Section II) that there is an $\mathcal{L}_{\mathbb{N}}$-theory (without T) that encodes a sufficient theory of syntax for theories of truth. For the sake of simplicity in the subsequent argument, let us fix any such $\mathcal{L}_{\mathbb{N}}$-theory M, e.g. PA. Hence, truth can be added, as a new logical device, to any $\mathcal{L}_{\mathbb{N}}$-theory B including M, which results in an $\mathcal{L}_{\mathbb{N}}^{+}$-theory of the form $\mathcal{E} + \mathsf{B}$ with B as its base theory. It is to be noted that $\mathcal{E} + \mathsf{B}$ can be an adequate theory of *truth* without embodying an adequate theory of *arithmetic*. Some might not accept B as an adequate theorisation of arithmetic for various reasons, but, as a logical device, truth should be neutral to one's view of what an adequate theorisation of arithmetic should be like. As long as B is consistent and can be seen as a theorisation of arithmetic, B is a logically coherent and possible theorisation of arithmetic, whether one likes it or not. Any theory (as a first-order theory including the logical axioms of first-order logic) is an adequate theorisation of first-order logical notions regardless of the adequacy of the theory as a theorisation of its subject. Similarly, $\mathcal{E} + \mathsf{B}$ should be seen as an adequate theorisation of truth *qua* a logical device applied to the particular theorisation B of arithmetic, regardless of the adequacy of B as a theorisation of arithmetic.

Given these preliminary considerations, let us draw the conservativeness requirement from the logicality thesis. Let B be any consistent $\mathcal{L}_{\mathbb{N}}$-theory including M. Suppose $\mathcal{T} + \mathsf{B}$ is not conservative over B for some set $\mathcal{T}$ of axioms essential to truth. Each axiom essential to truth is naturally counted in $\mathcal{E}$, and thus $\mathcal{E} + \mathsf{B}$ is not conservative over B. Take an $\mathcal{L}_{\mathbb{N}}$-sentence σ such that $\mathcal{E} + \mathsf{B} \vdash \sigma$ but $\mathsf{B} \nvdash \sigma$. If truth is logical, then $\mathcal{E} + \mathsf{B} \vdash \sigma$ means that σ is derivable by means of the non-logical axioms belonging to B and the logical axioms (and, possibly, rules) for truth and other logical notions, and thus σ

[10] Precisely speaking, $\mathcal{E}$ should be invariant across different subjects and their theorisations modulo a language *and a theorisation of the syntax of the target language of truth*. The exact formulation of $\mathcal{E}$ depends on the way the syntax of the target language is theorised. For example, we have different codings (or Gödel numberings) of syntactic objects in arithmetic, and the exact formulation of $\mathcal{E}$ varies depending on the choice of coding.

should be seen as a logical consequence of B, but $\mathsf{B} \nvdash \sigma$ states that σ is not a logical consequence of B. It also follows that $\mathsf{C} := \mathsf{B} + \neg\sigma$ is consistent, but $\mathcal{E} + \mathsf{C}$ is inconsistent, and, by the same reasoning, C is viewed as both logically consistent and inconsistent. Formally speaking, this is not a genuine contradiction because the non-conservativeness at issue only means that σ is a logical consequence of B in one 'logic' that includes truth as a logical notion, but not a logical consequence of B in another 'logic', our current first-order predicate logic, which does not include truth as a logical notion. However, the supposed non-conservativeness implies that the two logics are not identical. Hence, unless we reject and significantly revise our current concept of logic (and thus withdraw the second stipulation made in Section II), as Shapiro (1998) suggests on the basis of his conservativeness argument (see fn 4), the theory $\mathcal{E} + \mathsf{B}$ of truth ought to be conservative over its base theory B, and thus the conservativeness requirement is necessitated by the logicality thesis; we will consider the possibility of revising logic later in Section V. Let us call the conservativeness argument that employs the logicality thesis in justification of the conservativeness requirement the *conservativeness argument from logicality*.[11]

Now, we have two types of conservativeness arguments on independent grounds: the logicality thesis and the insubstantiality thesis each justify the conservativeness requirement without resorting to the other. This significantly enhances the efficacy of the conservativeness argument: if one successfully finds a non-conservative adequacy condition, then one can put considerable pressure on our deflationist to give up both the logicality and insubstantiality theses, the two core doctrines of deflationism, via the conservativeness argument.

Furthermore, the logicality thesis blocks one possible way out from a certain type of the conservativeness argument *from insubstantiality*. To see this, let us turn to another example of the conservativeness argument due to Shapiro (1998), in which he employs the following two adequacy conditions.

Compositionality requirement (*Comp*): An adequate theory of truth with any base theory ought to prove the following (typed) *compositional axioms*:

(C1) for each $\mathcal{L}_{\mathbb{N}}$-atomic formula $Rx_1 \ldots x_k$, it is true of objects $a_1, \ldots, a_k$, if and only if $Ra_1 \ldots a_k$;
(C2) for each $\mathcal{L}_{\mathbb{N}}$-sentence σ, $\neg\,\sigma$ is true, iff σ is not true;
(C3) For each $\mathcal{L}_{\mathbb{N}}$-sentences σ and τ, $\sigma \vee \tau$ is true, iff either σ or τ is true;

[11] Note that the (supposed) topic-neutral aspect of truth is not used in the conservativeness argument from logicality in itself, though it is used in laying out the formal setting adopted here and will be used to block one possible escape from the conservativeness argument below.

(C4) For each $\mathcal{L}_{\mathbb{N}}$-formulas $\varphi(v)$ with one free variable v, $\exists v\varphi(v)$ is true, iff φ is true of some object;

we will denote the set of the (typed) compositional axioms by $\mathcal{CT}$.[12]

Induction requirement (*Ind*): An adequate theory of truth with any base theory ought to prove all the instances of arithmetical induction in the language $\mathcal{L}^{+}_{\mathbb{N}}$:

$$(\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(x+1))) \rightarrow \forall x\varphi(x), \text{ for all } \mathcal{L}^{+}_{\mathbb{N}}\text{-formulas } \varphi(x);$$

we will denote the set of all these instances by $\mathcal{IN}$.[13]

Under these requirements *Comp* and *Ind*, the conservativeness argument goes on to conclude that there is no adequate deflationist theory of truth with PA as its base theory, since $\mathsf{PA} + \mathcal{CT} + \mathcal{IN}$ proves *Con*(PA). However, this version of the conservativeness argument cannot be applied to some other base theories; for example, the addition of $\mathcal{CT}$ and $\mathcal{IN}$ to the first-order part of the ramified analysis, namely, PA plus transfinite induction up to the Feferman–Schütte ordinal Γ_0, yields no new $\mathcal{L}_{\mathbb{N}}$-theorem. Now, without the logicality thesis, truth need not be universally applicable, and our deflationist might suggest that we should reconstrue the insubstantiality thesis as necessitating not the full-fledged conservativeness requirement for every base theory, but only a certain restriction of it to *some* base theories.[14] Thereby, they could circumvent the conservativeness argument from insubstantiality of this type, which employs *Comp* and *Ind*, by denying that PA (or any other theory falling prey to this type of the conservativeness argument) is an appropriate theory to which truth is applied. However, this way out is blocked by the logicality thesis, since truth should be applicable to any base theory B as a logical device, regardless of any qualification of B, as long as B can encode a sufficient theory of syntax,

[12] See Fujimoto (2021) for an argument in favor of *Comp*.

[13] There is controversy over *Ind*. On the one hand, Field (1999) rejects it because he views each instance of $\mathcal{IN}$ as postulated not by virtue of the deflationary concept of truth, but by consideration of 'something about our idea of natural numbers' and 'nothing about truth' (p. 539). On the other hand, some argue that the extension of arithmetical induction for $\mathcal{L}^{+}_{\mathbb{N}}$ is not required by one's particular choice of arithmetic as the base subject, but rather by one's more general commitment to the inductive structure of the bearers of truth, i.e., sentence types, and thus is essential to truth because the theory of syntax is necessary part of any adequate theory of truth; see Shapiro (2004) and Fujimoto (2019) for such arguments.

[14] The insubstantiality thesis seems to require that at least some adequate theories of truth (for each base subject) can be divided into the truth part $\mathcal{T}$ purely about truth and the truth-free base part B purely about the non-semantic base subject so that $\mathcal{T} + \mathsf{B}$ is conservative over B. For, otherwise, every adequate theory of truth would inevitably include some axioms that contain T but make some substantial contribution to the theorisation of the base subject, which means that a theorisation of truth always results in the addition of some axioms about the base subject and, at the same time, excludes the possibility of some other axioms about the base subject (i.e., their negations); this seems to indicates the substantiality of truth.

and truth should still be viewed as non-logical if it yields a new truth-free consequence.

IV. THE TOPIC-NEUTRALITY ARGUMENT

In the last section, we have seen that the logicality thesis provides a new ground for the conservativeness argument, and even strengthens a certain type of the conservativeness argument. In the present section, we will see that it also yields a completely new type of an argument in challenge to deflationism, which I call the *topic-neutrality argument*.

IV.1 The topic-neutrality requirement

I will first formulate a central premise for the topic-neutrality argument.

Recall that $\mathcal{E}$ is supposed to only (but adequately) axiomatise the essential function of truth as a logical device. As I have argued in Section III.2, given any $\mathcal{L}_{\mathbb{N}}$-theory B including the fixed minimal theory M, the $\mathcal{L}^{+}_{\mathbb{N}}$-theory $\mathcal{E}+\mathsf{B}$ ought to be an adequate theory of truth. Now, as before, there must be some conditions for an $\mathcal{L}^{+}_{\mathbb{N}}$-theory to be an adequate theory of truth. Such conditions might be *GEP*, *Disq*, *Comp*, *Ind*, or something else, but, no matter what they are, $\mathcal{E}$ should be solely responsible for the adequacy of the theory $\mathcal{E}+\mathsf{B}$ as a theory of truth. The axioms of a logical device should solely characterise the nature and function of the device adequately and should not depend on a theorisation of the subject of discourse to which it is applied or added. Hence, if truth is logical, $\mathcal{E}+\mathsf{B}$ should meet the adequacy conditions regardless of our choice of B.[15]

This consideration leads us to the following new requirement under the logicality thesis.

Topic-neutrality requirement: Truth should be applicable to any base theory in a uniform way so that $\mathcal{E}+\mathsf{B}$ satisfies the adequacy conditions for theories of truth for every base theory B.

The topic-neutrality argument argues that some natural adequacy conditions are incompatible with this topic-neutrality requirement.

[15] As I have argued in Section III.2, a base theory B can be any theory that embodies a sufficient theory of syntax and can be seen as a theorisation of the base subject, but it is, of course, difficult to precisely define what it means for a formal theory to be a theorisation of a given subject. However, for the purpose of the present paper, I believe that we need not worry about this issue. All my formal arguments still hold valid even if we restrict ourselves to $\mathcal{L}_{\mathbb{N}}$-theories that are widely seen as theorisations of arithmetic; more precisely, we can take such theories as B and C in the formal results in the present paper; see also fn 18 below.

Before proceeding further, let me point to one alternative possible account of topic-neutrality. Some philosophers propose to characterise topic-neutrality in terms of invariance under permutations; see McCarthy (1981) and Westerståhl (1985) for example. This is a derivative of a more general idea of characterising logicality in such terms, which is nowadays called the Tarski–Sher thesis after Tarski (1986) and Sher (1991). It is known that logicality in this sense goes far beyond our current concept of first-order predicate logic: McGee (1996) showed that an operation (on a fixed domain) is invariant under permutations if and only if it is definable in the full infinitary logic $L_{\infty,\infty}$. Hence, some authors, such as Feferman (1999) and Bonnay (2008), suggest characterising logicality in terms of invariance not under permutations, but under some weaker 'similarity' relations across structures.[16]

However, the idea of characterising logicality or topic-neutrality in terms of invariance, under permutations or some 'similarity' relation, does not fit well with our current framework stipulated in Section II. Above all, while we focus on the axiomatic approach to truth, the invariance characterisation is defined in terms of model theory. Apart from this obvious problem, there is a more general technical subtlety in applying the invariance criterion to truth. Recall that truth presupposes syntactic entities and a theory of syntax. Hence, truth cannot be treated as an operator on all arbitrary structures, but should be treated as an operator on structures of a certain special type that contain syntactic entities (or their codes) and embody a sufficiently rich inner structure of these entities. Therefore, we have to specify an appropriate class $\mathcal{M}$ of such structures on which truth is to be defined as an operator, and the relevant 'similarity' relation $\sim$, against which invariance is tested, has to be restricted to such an $\mathcal{M}$. This already deviates from the standard formal framework for the invariance approach to logicality. Furthermore, we have to choose, among many options, an appropriate formal characterisation, say, K, of truth as an operator on those structures. Choosing an appropriate triple of such an $\mathcal{M}$, $\sim$, and K is far from an obvious task, and there still remains a lot of work to be done toward the formal implementation of the idea of applying the invariance criterion to truth.[17] Hence, while it is worth pursuing this alternative approach and comparing it with my own, I leave this task for another occasion.

[16] There are subtle technical differences between the formulations of the idea by these authors, among which Bonnay's (2008) approach is the most general and subsumes the others; for more details, further references, and the historical matters, see Feferman (1999) and Bonnay (2008).

[17] Firstly, the choice of $\mathcal{M}$ may depend on the choice of K. Secondly, it may not be possible to treat truth as an operator defined in a unique and uniform manner across structures from $\mathcal{M}$. Thirdly, choosing a right 'similarity' relation is always a considerable challenge to any attempt to characterise logicality in terms of invariance. Bonnay and Galinon (2018) try to overcome these difficulties by treating truth as an operator on a certain extension of each structure, called an *alethic extension*, by means of which truth is made definable in a unique and uniform way across all arbitrary structures, and showed the invariance of the thus characterised notion of truth under many reasonable 'similarity' relations. However, an alethic extension of a structure, say, M, is obtained, in essence, by attaching the standard model of arithmetic as the domain of a new sort, then treating the standard model as the structure of syntax, and thereby uniquely defining truth

IV.2 The topic-neutrality argument

The topic-neutrality argument depends on the adequacy condition one imposes upon theories of truth. Hence, it is an umbrella term encompassing arguments of a certain type with the same structure. I will first present a sample of the topic-neutrality argument that employs *GEP* and *Disq* as adequacy conditions for theories of truth, which provides a template for the topic-neutrality argument.

Proposition 1. *Let* B *and* $\mathcal{E}$ *be recursive sets of* $\mathcal{L}_{\mathbb{N}}$*- and* $\mathcal{L}_{\mathbb{N}}^{+}$*-sentences, respectively, such that* $\mathsf{B} \vdash \mathsf{M}$*. If* $\mathcal{E} + \mathsf{B}$ *is consistent, then there is a primitive recursive consistent set* C *of* $\mathcal{L}_{\mathbb{N}}$*-sentences such that* $\mathsf{C} \vdash \mathsf{B}$ *and* $\mathcal{E} + \mathsf{C}$ *have the same* $\mathcal{L}_{\mathbb{N}}$*-theorems as* C*.*

Proof. Let X be the set of the $\mathcal{L}_{\mathbb{N}}$-sentences that are provable in $\mathcal{E} + \mathsf{B}$. Since X is recursively enumerable, it follows by Craig's trick (Craig , 1953) that X is axiomatisable by some primitive recursive set C of $\mathcal{L}_{\mathbb{N}}$-sentences. Now, suppose $\mathcal{E} + \mathsf{C} \vdash \sigma$ for an $\mathcal{L}_{\mathbb{N}}$-sentence σ. Since C axiomatises X, we have $\mathcal{E} + \mathsf{B} \vdash \sigma$ and thus $\mathsf{C} \vdash \sigma$. □

Corollary 2. *Let* B *and* $\mathcal{E}$ *be as above. If* $\mathcal{E} + \mathsf{B}$ *is consistent with* $\mathcal{TB}$ *(see page 7 for its definition), then there is a primitive recursive consistent set* C *of* $\mathcal{L}_{\mathbb{N}}$*-sentences such that* $\mathsf{C} \vdash \mathsf{B}$ *and* $\mathcal{E} + \mathsf{C} \nvdash$ *GRef*(C)*.*

Proof. Let $\mathcal{E}' := \mathcal{E} + \mathcal{TB}$. There is a primitive recursive consistent set C of $\mathcal{L}_{\mathbb{N}}$-sentences such that $\mathsf{C} \vdash \mathsf{B}$ and $\mathcal{E}' + \mathsf{C}$ is conservative over C. If $\mathcal{E} + \mathsf{C} \vdash$ *GRef*(C), then $\mathcal{E}' + \mathsf{C} \vdash$ *Con*(C) and thus $\mathsf{C} \vdash$ *Con*(C), which is impossible due to Gödel's second incompleteness theorem. □

Hence, no matter what $\mathcal{E}$ is given, we can always find a base theory C such that $\mathcal{E} + \mathsf{C}$ fails to meet both *GEP* and *Disq* at the same time, which conflicts with the topic-neutrality requirement, under the assumption that *GEP* and *Disq* are necessary conditions for adequate theories of truth.[18]

as a predicate on the domain of the new sort whose extension consists exactly of the standard names of the sentences true in M. Hence, their formal setting could be seen as a strengthened, model-theoretic version of Leigh and Nicolai's (2013) and conflicts with our fourth stipulation made in Section II, and, more importantly, their results crucially rely on the use of the standard model of arithmetic, which is not acceptable from the viewpoint of the axiomatic approach to truth.

[18] One might be concerned that such a 'counterexample' theory C constructed by Craig's trick may be too ad hoc and unnatural to be counted as a base theory of truth (cf. fn 15). However, I use Craig's trick only for making the results as general as possible, and, in almost every case considered in the literature (of which I am aware), we can find a natural, independently motivated theory B such that B itself can be taken as C in Corollary 2 and Propositions 3 and 4. For example, if $\mathcal{E}$ or $\mathcal{T}$ there is taken to be $\mathcal{CT}$ plus $\mathcal{IN}$, then we can take the theory of autonomous progression of uniform reflection principles as such $\mathsf{B} = \mathsf{C}$; if it is $\mathcal{CT}$ plus $\mathcal{IN}$ restricted to Δ_0-formulas

In general, the topic-neutrality argument has the following structure:

(L1) the topic-neutrality requirement must be met;
(L2) there is some adequacy condition for theories of truth and some base theory C such that the condition is not met by $\mathcal{E} + \mathsf{C}$;
(L3) therefore, the logicality thesis and thus deflationism about truth are untenable.

The success of the topic-neutrality argument, of course, depends on the adequacy condition one employs in the second step (L2). We have seen that if one employs *GEP* and *Disq*, then the topic-neutrality argument achieves its aim. In contrast, if one adopts *Comp* and *Ind* instead, then it does not, since both $\mathcal{CT}$ and $\mathcal{IN}$ are fixed sets of $\mathcal{L}^{+}_{\mathbb{N}}$-sentences, and thus *Comp* and *Ind* are trivially met by any such $\mathcal{E}$ that includes $\mathcal{CT}$ and $\mathcal{IN}$.

IV.3 Comparison with the conservativeness argument

The conservativeness argument and the topic-neutrality argument are different arguments. The obvious difference between them is that the former aims to conclude the untenability of deflationism from the *derivability* of something, but the latter aims to draw the same conclusion from the *underivability* of something.[19] However, beyond such a difference in form, there is more substantial differences in their applications.

Firstly, the topic-neutrality argument even challenges some *conservative* adequacy conditions. For example, given a base theory B, let $\mathit{Tr}(\mathsf{B})$ denote the $\mathcal{L}^{+}_{\mathbb{N}}$-sentence expressing that all the *axioms* of B are true; note that, in contrast to $\mathit{GRef}(\mathsf{B})$, $\mathit{Tr}(\mathsf{B})$ only asserts the truth of the axioms of B and says nothing about other logical consequences of B. Now, let us consider the following weaker version of *GEP*, in terms of $\mathit{Tr}(\mathsf{B})$ instead of $\mathit{GRef}(\mathsf{B})$.

Weak generalised equivalence principle (*WGEP*): An adequate theory of truth with the base theory B ought to prove $\mathit{Tr}(\mathsf{B})$.

It is well-known folklore that $\mathsf{B} + \mathcal{CT} + \mathit{Tr}(\mathsf{B})$ is always conservative over any B (including M),[20] and so is $\mathsf{B} + \mathcal{TB} + \mathit{Tr}(\mathsf{B})$ (since $\mathsf{B} + \mathcal{CT} \vdash \mathcal{TB}$). Hence,

(see Wcisło and Łełyk 2017), then we can choose the theory of finitely iterated uniform reflection principles as such; if it consists of the truth axioms of the Kripke–Feferman theory (KF), then we can take the arithmetical part of ramified analysis as such (though a slight tweaking is required to adapt Leigh and Nicolai's (2013) formal framework for KF).

[19] One might suspect that this indicates a conceptual incompatibility of the topic-neutrality and conservativeness requirements, but I don't think so. The former requires the derivability of some sentences that are supposed to certify the adequacy of the theory of truth in question and thus naturally expected to be sentences involving T; in contrast, the latter requires the underivability of some sentences that are expressible without T and determined by the base theory alone.

[20] See Wcisło and Łełyk (2017, n. 6) for more details of this folklore.

WGEP is a harmless condition from the point of view of the conservativeness requirement, even when it is combined with *Comp*. However, as the next proposition shows, it conflicts with the topic-neutrality requirement.

Proposition 3. *Let* B *and* $\mathcal{E}$ *be as in Proposition 1. If* $\mathcal{E}$ *is consistent with* $\mathsf{B} + \mathcal{CT} + \mathcal{IN}$*, then there is a primitive recursive consistent set* C *of* $\mathcal{L}_{\mathbb{N}}$*-sentences such that* $\mathsf{C} \vdash \mathsf{B}$ *and* $\mathcal{E} + \mathsf{C} \nvdash \mathit{Tr}(\mathsf{C})$.[21]

Proof. As before, the set of $\mathcal{L}_{\mathbb{N}}$-sentences provable in $\mathcal{E} + \mathsf{B} + \mathcal{CT} + \mathcal{IN}$ can be axiomatised by a primitive recursive set C of $\mathcal{L}_{\mathbb{N}}$-sentences. By the same argument as Proposition 1, we can show that $\mathcal{E} + \mathsf{C} + \mathcal{CT} + \mathcal{IN}$ is conservative over C. Now, $\mathcal{CT} + \mathcal{IN} + \mathsf{B}$ proves that first-order logic preserves truth and, therefore, that for every primitive recursive set $\mathcal{Y}$ of $\mathcal{L}_{\mathbb{N}}$-sentences, $\mathit{Tr}(\mathcal{Y})$ implies $\mathit{Con}(\mathcal{Y})$; in particular, it proves that $\mathit{Tr}(\mathsf{C})$ implies $\mathit{Con}(\mathsf{C})$. Suppose for contradiction that $\mathcal{E} + \mathsf{C} \vdash \mathit{Tr}(\mathsf{C})$. Then, $\mathcal{E} + \mathsf{C} + \mathcal{CT} + \mathcal{IN}$ proves $\mathit{Tr}(\mathsf{C})$ and thus $\mathit{Con}(\mathsf{C})$, which implies $\mathsf{C} \vdash \mathit{Con}(\mathsf{C})$; a contradiction. □

Next, we will see that some effective deflationist rejoinders to the conservativeness argument fail to save deflationism from the topic-neutrality argument. I will give two examples of such below.

Firstly, it is a widely shared view that the mathematical structure that the theory of syntax describes is essentially the same as that of natural numbers. Hence, when a theory of truth is formulated with an arithmetical coding of the theory of syntax within an arithmetical base theory B, B is used as a theory of the non-semantic base subject and a theory of syntax at the same time, and the non-semantic base content and the syntactic content of B exactly coincide. Taking this 'entanglement' of the two roles of an arithmetical base theory B into account, even if $\mathcal{E} + \mathsf{B}$ is not conservative over B, the failure of conservation can be interpreted to only mean that truth is not conservative over a theory of syntax, but this is not a problem for deflationism, because truth axioms and a theory of syntax always come in one package and it makes little sense to separate and compare them in terms of conservativeness.[22] However, this line of defense of deflationism against the conservativeness argument is of no help against the topic-neutrality argument, since it provides no excuse for the *underivability* of wanted consequences.

The second example is the so-called theories of truth with a 'disentangled' theory of syntax. The aforementioned 'entanglement' of the two roles in a single arithmetical base theory is an inevitable consequence of the customary formal setting for theories of truth, in which the theory of bearers of truth is assumed to be embedded in the theory of the base subject. Having reflected upon this point, Heck (2009) and Leigh and Nicolai (2013) proposed a new

[21] We can strengthen the statement, without changing the proof, by replacing $\mathcal{IN}$ in the premise with the restriction of $\mathcal{IN}$ to Δ_0-formulas containing the truth predicate (see fn 18).

[22] See Shapiro (2004) and Fujimoto (2019) for this line of argument.

type of theory of truth in which a theory of syntax is given as a completely separate theory from the base theory B, with a new domain of its own objects separate from the domain of the non-semantic objects of B. The language of this type of theory is a three-sorted first-order language with the first sort o for the objects of the base subject, the second sort s for the syntactic objects as the bearers of truth, and the third sort sq for sequences of objects of the first sort as satisfaction sequences (or variable assignments); then, it has a special predicate $Sat(x, y)$ of type $sq \times s$, expressing that 'the formula y is satisfied by a sequence x, which takes an object of the third sort sq as its first argument and an object of the second sort s as its second argument. A theory of truth (of arithmetic) of this type is thereby defined as the conglomeration of the following four theories:

- an $\mathcal{L}_{\mathbb{N}}$-theory B as a theory of the first sort o;
- a theory Syn of syntax of the second sort s, which is usually taken to be some standard arithmetical theory such as PA;
- a theory Sq of sequences of the third sort sq;
- a theory $\mathcal{T}$ of the satisfaction predicate;

the subtle difference between the satisfaction predicate and the truth predicate can be ignored in the current context, and we here identify theories of satisfaction with those of truth. In this three-sorted language, the consistency statement is naturally formulated as a statement about objects not of the first sort o, but of the second sort s, and so are assertions of the truth of sentences about the base subject, such as the global reflection principle in particular. The most important consequence of this new formal setting, in the current context, is that the global reflection principle for the base theory B in terms of the second sort s does not involve any mention of objects of the base subject (i.e., of the sort o) and the consistency statement of B is no longer a statement in the language of B; in general, the theory of truth of this type is always conservative over B (as a theory of the sort o).[23] Hence, by construing *GEP* as requiring the provability of the global reflection principle in terms of s, a theory of truth of this type may derive the global reflection principle without breaking the conservativeness and thereby make *GEP* (as well as *Disq*) compatible with the conservativeness requirement. However, as the next proposition shows, it is still susceptible to the topic-neutrality argument.

Proposition 4. *Let* $\mathsf{T} := \mathcal{T} + \mathsf{Syn} + \mathsf{Sq} + \mathsf{B}$ *be a recursive theory of truth with a 'disentangled' theory of syntax. Take a canonical translation* $\mathcal{I}$ *of the three-sorted language of* T *into* $\mathcal{L}^{+}_{\mathbb{N}}$*, in which syntactic objects of the sort s and sequences of the sort sq are translated into their arithmetisations, and the base objects of the sort o are translated verbatim. If* $\mathcal{I}(\mathsf{T})$ *is consistent, then there is a primitive recursive consistent set* C *of* $\mathcal{L}_{\mathbb{N}}$*-sentences with* $\mathsf{C} \vdash \mathsf{B}$

[23] See Halbach (2010, Section 22.2) or Leigh and Nicolai (2013, Corollary 3.12).

such that $\mathcal{T} + \mathsf{Syn} + \mathsf{Sq} + \mathsf{C}$ *does not prove the consistency of* C*, nor the global reflection principle for* C*, even in terms of the second sort* s*.*[24]

Proof. Take a primitive recursive axiomatisation C of the $\mathcal{L}_{\mathbb{N}}$-theorems of $\mathcal{I}(\mathsf{T})$. Let T' denote $\mathcal{T} + \mathsf{Syn} + \mathsf{Sq} + \mathsf{C}$. As before, if $\mathcal{I}(\mathsf{T}') \vdash \sigma$, then $\mathcal{I}(\mathsf{T}) \vdash \sigma$ and thus $\mathsf{C} \vdash \sigma$. Suppose for contradiction that T' proves the consistency statement of C in terms of the sort s. Then, $\mathcal{I}(\mathsf{T}')$ proves its $\mathcal{I}$-translation, namely, the standard $\mathcal{L}_{\mathbb{N}}$-expression *Con*(C) of the consistency of C, which implies $\mathsf{C} \vdash$ *Con*(C); a contradiction. □

V. POSSIBLE SOLUTIONS

In this final section, I will discuss how our deflationist could respond to the conservativeness argument from logicality and the topic-neutrality argument.

The most straightforward solution to the topic-neutrality argument is to deny the adequacy conditions employed in the premise (L2). In response to the particular topic-neutrality argument employing *GEP* or *WGEP*, deflationists could simply deny the requirement for the derivability of *GRef*(B) or *Tr*(B). This strategy is equally effective as a solution to the conservativeness argument, either from insubstantiality or logicality. For example, Azzouni (1999) claims, in response to Shapiro's (first) conservativeness argument, that '[w]hat *is* true (and how) is not, properly speaking, part of *the* theory of truth' (542) and,

> the capacity to establish (nonlogical!) truths and generalizations about such goes quite beyond what a first-order deflationist calls a deflationist theory of truth—and this regardless of how obvious such truths and generalizations (about them) are. (542)

Azzouni's rejoinder sounds sensible on its own right, but Shapiro and Ketland have their own points as well. In particular, from the perspective of the logicality thesis, it can be objected against Azzouni that 'what is true (and how)' is not part of *logic*, either, but logic still makes conjunctions of finitely many non-logical axioms of a theory, as well as each individual non-logical axiom, logical consequences of the theory, and so if truth is part of logic, then it seems reasonable to require that conjunctions of (possibly) infinitely many non-logical axioms of a theory B, such as *Tr*(B), should be made

[24] The consistency of $\mathcal{I}(\mathsf{T})$ is not always a natural assumption even if T is consistent. However, when B is a 'standard' theory of arithmetic, such as PA, we have a reasonable ground for assuming it. If Syn and the theory of syntax encoded in B represent the same (or sufficiently similar) structure of syntactic objects, then the two theories of syntax are expected to be adequately correlated so that we can freely pass from one theory of syntax to the other; consider, for example, the case where $\mathsf{B} = \mathsf{Syn} = \mathsf{PA}$ or B is 'maximally rich' in the sense of Fujimoto (2019). This intuition is formally expressed by the 'bridge laws' of Leigh and Nicolai (2013, Section 3.4), which postulates an isomorphism between the structures represented by the two theories of syntax, and if the 'bridge laws' are consistent with T, then $\mathcal{I}(\mathsf{T})$ is consistent.

logical consequences of B by virtue of truth as a logical device of infinite conjunction.

After all, whether this 'straightforward' solution succeeds or not boils down to the question as to what is an adequacy condition for a theory of truth. The debate over the adequacy condition for theories of truth is far from settled, and I do not intend to settle this long-lasting question here nor assess the success of the solution at stake.[25] My primary aim in the present paper is to suggest a new type of argument that poses a new type of threat to deflationism and to add one possibly useful tool to the anti-deflationists' toolbox.

In concluding the present paper, I will then discuss how deflationists could respond to the topic-neutrality argument and the conservativeness argument from logicality even if the conditions *GEP*, *WGEP*, etc., are accepted.

In the formal setting considered in the present paper, we start with a recursive set B of $\mathcal{L}_{\mathbb{N}}$-sentences, as the base theory of a theory of truth. Then, the topic-neutrality argument blames the theory $\mathcal{E} + \mathsf{B}$ of truth for not logically implying something, and the conservativeness argument from logicality blames it for logically implying something that does not logically follow from B alone. Both arguments concern the logical consequences of $\mathcal{E} + \mathsf{B}$ and/or B. However, it is ordinary first-order logic here that determines what are logical consequences of these theories. Recall that the conservativeness requirement is necessitated by the logicality thesis *unless* we reject and significantly revise our current concept of logic. Hence, one possible solution is to reject the current first-order logic and adopt an extended notion of logical consequence with truth as a new logical notion. Now, suppose $\mathcal{E} + \mathsf{B}$ is not conservative over B. There is an $\mathcal{L}_{\mathbb{N}}$-sentence σ such that $\mathcal{E} + \mathsf{B} \vdash \sigma$ but $\mathsf{B} \nvdash \sigma$ and that $\mathcal{E} + (\mathsf{B} + \neg\sigma)$ is inconsistent but $\mathsf{B} + \neg\sigma$ is consistent. With this 'logical revisionist' solution, one can construe this situation as only meaning that σ *is* a logical consequence of B in the logical revisionists' sense, and, similarly, $\mathsf{B} + \neg\sigma$ is already logically *in*consistent in their sense; hence, the conservativeness requirement is not necessitated by the logicality thesis, and thus the logicality of truth is by no means undermined by the supposed non-conservativeness.[26]

This 'logical revisionist' solution also nicely deals with the topic-neutrality argument. Given a set B of axioms of the base subject, the ordinary first-order definition of logical consequence prescribes every member of B to be a logical consequence of B. Hence, when the notion of logical consequence is extended

[25] My own view is that the solution at stake is not very promising: I recently proposed in Fujimoto (2021) a new adequacy condition for theories of truth that raises another type of the conservativeness argument.

[26] The label 'logical revisionism' is usually used to express an idea of changing the meaning, axioms, and/or rules of the existing logical vocabulary, particularly, to denote a stance advocating a non-classical logic weaker than classical logic. Here, I rather mean thereby a stance advocating an extension of logic with new logical vocabulary without making any change to the original logical vocabulary (which is, though, a 'revision' of logic anyway).

with truth, it seems sensible to stipulate that the infinite conjunction of all members of B, i.e., $\mathit{Tr}(\mathsf{B})$, is also a logical consequence of B in the extended sense, even though $\mathit{Tr}(\mathsf{B})$ is not a logical consequence of B in the ordinary first-order sense. This suggests a revision of the definition of logical consequence so that each member of $\mathcal{E}$ is treated as a new logical axiom, and $\mathit{Tr}(\mathsf{B})$, as well as each logical consequence of $\mathcal{E} + \mathsf{B}$ in the ordinary first-order sense, is stipulated to be a logical consequence of B in the extended sense. Furthermore, depending on $\mathcal{E}$, more 'substantial' consequences may logically follow from B in the extended sense; for example, if $\mathcal{E} + \mathsf{B}$ satisfies *Comp* and *Ind*, then $\mathit{GRef}(\mathsf{B})$ becomes a logical consequence of B in the extended sense, since $\mathcal{CT}$, $\mathcal{IN}$, and $\mathit{Tr}(\mathsf{B})$ jointly imply $\mathit{GRef}(\mathsf{B})$ (in ordinary first-order logic). This 'logical revisionist' scenario seems to well capture Ketland's intuition behind his postulation of *GEP*, and resolves the problem posed by the topic-neutrality argument employing *GEP* or *WGEP*.

However, the 'logical revisionist' solution seems to still require deflationists to give up the insubstantiality thesis (so interpreted as to incur the conservativeness requirement), while it allows them to maintain the logicality thesis. Whether truth is logical or not, and no matter how the notion of logical consequence is extended via the addition of truth, if there is an $\mathcal{L}_{\mathbb{N}}$-sentence σ such that $\mathcal{E} + \mathsf{B} \vdash \sigma$ and $\mathsf{B} \nvdash \sigma$, then truth is anyway construed as bringing about new knowledge of the base subject, and this is naturally seen as in conflict with the insubstantiality thesis. Namely, if we stick to the insubstantiality thesis, then the 'logical revisionist' solution does not save deflationism from the conservativeness argument *from insubstantiality*. Hence, in order to adopt the solution, the logicality and insubstantiality theses should be taken to be independent of each other, and then only the latter should be abandoned.

One might be concerned that the insubstantiality thesis is not independent of the logicality thesis because anything logical should be insubstantial and thus the latter implies the former. However, it has been argued by several authors that a logical notion need not be insubstantial in such a way that necessitates 'conservativeness'. For example, Galinon (2015) points out that the logical notion of negation $\neg$ brings about new negation-free consequences that cannot be derived without negation, such as negation-free instances of Peirce's law, namely, $((\varphi \to \psi) \to \varphi) \to \varphi$ for negation-free formulas φ and ψ.[27] Take any standard deductive system $\mathcal{D}$ of classical first-order logic. The negation-free fragment $\mathcal{D}^-$ of $\mathcal{D}$ is obtained by completely removing negation and its axioms and/or rules from $\mathcal{D}$. Then, each negation-free instance of Peirce's law is not derivable in $\mathcal{D}^-$, while it is a tautology and thus derivable in $\mathcal{D}$. That is to say, the addition of negation to $\mathcal{D}^-$ yields new consequences in a language without negation. This fact could be expressed as the non-conservativeness of

[27] See Galinon (2015) for more such examples. The same point is also made by Horsten (2009) and Picollo and Schindler (2018).

negation over the other first-order logical devices, but we would not conclude from this that negation is not logical. This seems to indicate that we actually do not require a logical device to be insubstantial in the sense at issue. Logic must be taken as an entirety of all its constituent logical devices, and thus the negation-free instances of Peirce's law are logically valid in one logic with negation, but not in another logic without negation. *We* take these instances as logically valid because *we* adopt the former logic, i.e., classical first-order logic, and the underivability of them in $\mathcal{D}^-$ implies nothing about the logical validity of them in the sense of *our* logic. Hence, the non-conservativeness at issue raises no problem with the logicality of negation; recall that it was for exactly the same reason that the logicality thesis does not necessitate the conservativeness requirement from the perspective of the 'logical revisionist' solution.

It goes beyond the scope of the present paper to answer whether the 'logical revisionist' solution is a correct one. To draw a decisive conclusion, we first need to settle what are the axioms essential to truth (as new logical axioms) and what the extended notion of logical consequence with truth should be like. Also, since this solution requires a drastic revision of our current concept of logic, it demands a holistic assessment of its impacts on a broad range of subjects.[28]

Another solution is to give up both the logicality and insubstantiality theses. However, what would be left for deflationism about truth after abandoning its two core doctrines? If our deflationists give up both of them, they ought to explain what kind of special trait is left to truth that makes it 'deflationary'. In particular, there is a plethora of non-logical, mathematically substantial resources that can achieve the expressive roles that truth is supposed to undertake, such as sets and full-fledged second-order quantification. If our deflationists give up both the logicality and insubstantiality of truth, how could they differentiate truth from these resources?

A more moderate solution is to deny that the insubstantiality thesis necessitates the conservativeness requirement or that the logicality thesis necessitates the conservativeness requirement and/or the topic-neutrality requirement; the rejection of the former has already been proposed by several authors (see fn 7). This solution demands that deflationists provide a precise characterisation of the insubstantiality and/or logicality (or 'quasi-logicality', 'broad logicality', etc.) of truth and then demonstrate that the thus characterised logicality or insubstantiality does not necessitate the requirement(s) at stake while still making truth 'deflationary'.

[28] Some deflationists call truth 'quasi-logical' (or 'broadly logical') instead of 'logical'; see fns 1 and 3. It is a possibility that deflationism can be saved from the predicament at issue by attributing 'quasi-logicality' to truth instead of logicality, but it remains to be further clarified and discussed what 'quasi-logicality' is, how it differs from logicality, and, more importantly, how and in what sense it still makes truth 'deflationary' while resolving the challenges posed by the conservativeness and topic-neutrality arguments.

I reserve my verdict as to whether the logicality thesis or deflationism is untenable. A number of issues remain to be settled in order to arrive at a final conclusion. Nonetheless, I believe that this article raises new issues to be seriously considered in the assessment of deflationism.[29]

REFERENCES

Azzouni, J. (1999) 'Comments on Shapiro', *The Journal of Philosophy*, 96: 541–44.

Bar-On, D., Horisk, C. and Lycan, W. G. (2000) 'Deflationism, Meaning and Truth-Conditions', *Philosophical Studies*, 101: 1–28.

Bar-On, D. and Simmons, K. (2007) "The Use of Force Against Deflationism: Assertion and Truth", in D. Greimann and G. Siegwart (eds) *Truth and Speech Acts: Studies in the Philosophy of Language*, 61–89. New York: Routledge.

Bonnay, D. (2008) 'Logicality and Invariance', *The Bulletin of Symbolic Logic*, 14: 29–68.

Bonnay, D. and Galinon, H. (2018) "Deflationary Truth Is a Logical Notion", in M. Piazza and G. Pulcini (eds) *Truth, Existence, and Explanation*, 71–88. Cham, Switzerland: Springer.

Cieśliński, C. (2015) 'The Innocence of Truth', *Dialectica*, 69: 61–85.

Craig, W. (1953) 'On Axiomatizability Within a System', *The Journal of Symbolic Logic*, 18: 30–2.

Damnjanovic, N. (2005) 'Deflationism and the Success Argument', *The Philosophical Quarterly*, 55: 53–67.

——— (2010) 'New Wave Deflationism', in C. D. Wright and N. J. L. L. Pedersen (eds) *New Waves in Truth*, 53–67. New York: Palgrave Macmillan.

Edwards, D. (2013) 'Truth as a Substantive Property', *Australasian Journal of Philosophy*, 91: 279–94.

Feferman, S. (1999) 'Logic, Logics and Logicism', *Notre Dame Journal of Formal Logic*, 40: 31–54.

Field, H. (1994) 'Deflationist Views of Meaning and Content', *Mind*, 103: 247–85.

——— (1999) 'Deflating the Conservativeness Argument', *The Journal of Philosophy*, 96: 533–40.

Fischer, M. (2015) 'Deflationism and Instrumentalism', in T. Achourioti, H. Galinon, J. Martínez-Fernández and K. Fujimoto (eds) *Unifying the Philosophy of Truth: Logic, Epistemology, and the Unity of Science*, vol. 36. Berlin: Springer.

Fujimoto, K. (2012) 'Classes and Truths in Set Theory', *Annals of Pure and Applied Logic*, 163: 1484–523.

——— (2019) 'Deflationism beyond Arithmetic', *Synthese*, 196: 1045–69.

——— (2021) 'The Function of Truth and the Conservativeness Argument', *Mind*, doi: 10.1093/mind/fzaa083.

Galinon, H. (2015) 'Deflationary Truth: Conservativity or Logicality?', *The Philosophical Quarterly*, 65: 268–74.

Halbach, V. (2001) 'How Innocent Is Deflationism?', *Synthese*, 126: 167–94.

——— (2010) *Axiomatic Theories of Truth*. Cambridge: CUP.

Heck, R. (2009) 'The Strength of Truth Theories', Unpublished manuscript. Available from: http://rkheck.frege.org/pdf/unpublished/StrengthOfTruthTheories.pdf. Accessed 10 January 2022.

Hill, C. S. (2002) *Thought and World: An Austere Portrayal of Truth, Reference, and Semantic Correspondence*. Cambridge: CUP.

Horsten, L. (1995) 'The Semantical Paradoxes, the Neutrality of Truth, and the Neutrality of the Minimalist Theory of Truth', in P. Cortois (ed.) *The Many Problems of Realism*, 173–87. Oxford: Tilburg University Press.

——— (2009) 'Levity', *Mind*, 118: 555–81.

——— (2011) *The Tarskian Turn*. Cambridge, Massachusetts: MIT Press.

[29] I am very grateful to the two anonymous referees for their helpful comments and constructive suggestions, which led to a considerable improvement of this article. I would also like to thank Leon Horsten for valuable comments on an earlier version of this article.

Horwich, P. (1998) *Truth*. Oxford: OUP.
——— (2010) *Truth–Meaning–Reality*. Oxford: OUP.
Ketland, J. (1999) 'Deflationism and Tarski's Paradise', *Mind*, 108: 69–94.
Künne, W. (2003) *Conceptions of Truth*. Oxford: OUP.
Leigh, G. and Nicolai, C. (2013) 'Axiomatic Truth, Syntax and Metatheoretic Reasoning', *Review of Symbolic Logic*, 6: 613–36.
Leitgeb, H. (2007) 'What Theories of Truth Should Be Like (but Cannot Be)', *Philosophy Compass*, 2: 276–90.
McCarthy, T. (1981) 'The Idea of a Logical Constant', *The Journal of Philosophy*, 78: 499–523.
McGee, V. (1996) 'Logical Operations', *Journal of Philosophical Logic*, 25: 567–80.
McGinn, C. (2000) *Logical Properties: Identity, Existence, Predication, Necessity, Truth*. Oxford: OUP.
Picollo, L. and Schindler, T. (2018) 'Deflationism and the Function of Truth', *Philosophical Perspectives*, 32: 326–51.
Shapiro, S. (1998) 'Proof and Truth: Through Thick and Thin', *The Journal of Philosophy*, 95: 493–521.
Shapiro, S. (2004) 'Deflation and Conservation', in V. Halbach and L. Horsten (eds) *Principles of Truth*, 2nd edn, 103–28. Frankfurt: Ontos Verlag.
Sher, G. (1991) *The Bounds of Logic*. Cambridge, Massachusetts: MIT Press.
Strollo, A. (2013) 'Deflationism and the Invisible Power of Truth', *Dialectica*, 67: 521–43.
Tarski, A. (1986) 'What are Logical Notions?', *History and Philosophy of Logic*, 7: 143–54.
Waxman, D. (2017) 'Deflationism, Arithmetic, and the Argument from Conservativeness', *Mind*, 126: 429–63.
Wcisło, B. and Łełyk, M. (2017) 'Notes on Bounded Induction for the Compositional Truth Predicate', *The Review of Symbolic Logic*, 10: 455–80.
Westerståhl, D. (1985) 'Logical Constants in Quantifier Languages', *Linguistics and Philosophy*, 8: 387–413.
Wyatt, J. (2016) 'The Many (yet Few) Faces of Deflationism', *The Philosophical Quarterly*, 66: 362–82.

University of Bristol, UK